I spoke with no-one concerning this exam except for a faculty member , 3768.

STATS 2019 Qualify Exam

Pre-hospital Diagnosis in Stroke

Jing Liao 65763768

Abstract

Background There is an immense interest in pre-hospital diagnostic of stroke in order to enable rapid acquisition in the acute care setting upon arrival. Our study focus on pre-hospital stroke diagnosis given clinical and EEG variables.

Method I used generalized linear model (GLM) by R software to analyze the collected data of 100 patients at a local hospital with clinical and EEG variables. The Likelihood ratio statistics is used to test the significance of the association of RACE and stroke given all other clinical variables in the model and the association varies by age.

For the predictive model, I applied LASSO to control the model complexity and selected the predictors. Then criteria used for further model selection is AIC. To avoid over-fitting, I firstly consider linear form of variables. After determining the subset of variables, model diagnostics would examine such assumption. At last, a goodness of fit test is performed to check whether the model is adequate. Model assessment through method of receiver operating caracteristic (ROC) and area under the curve (AUC) will also be presented and compared with model(only included clinical variables).

Results There is association between RACE and stroke given all other clinical variables.(P value$< 10^{-3}$). Patients with higher RACE are more likely to be diagnosed as stroke. However, there is no statistically significant evidence of a different association of RACE and stroke across the age.(P value$= 0.06$). The association between RACE and stroke is roughly the same regardless of age. The AUCs of predictive model are 0.213 and 0.217 higher than the simpler model with only clinical variables applied to training set and test set.

1 Introduction

Nowadays,[1][2] stroke is the 2nd leading cause of death in the world, the 4th leading cause of death in the United States. Approximately 800,000 strokes occur each year, every 40 seconds, someone has a stroke in US, every 4 minutes, someone dies of a stroke in the US.

There is an immense interest in pre-hospital of stroke in order to enable rapid acquisition in the acute care setting upon arrival. Currently, the standard clinical practice is to use Rapid Arterial Occlusion Evaluation (RACE) score to decide whether a patient had stroke or not.

Overwhelming evidence[7][8] has shown that some additional clinical variables along with the Electroencephalogram (EEG) changes immediately after brain ischemia can be used to identify patients with acute stroke.

Therefore, it would be our interest to explore the factors that associated with stroke and the accuracy in diagnostic of stroke based on clinical variables and EEG variables.

2 Methods

2.1 Data Collection

The data was collected on 100 patients who were admitted to the Emergency Department at a local hospital. The clinical variables considered in our study are RACE score, Last Known Well(LKW) time in hours, Gender, and Age. Those with suspected stroke underwent a 3 min EEG using a wireless dry-electrode system. We measured 100 EGG signals, which are denoted as E1 to E100, are summery values of 100 different electrodes over time. The outcome of interest is a binary variable called Stroke, which is set to 1 if the patient did in fact had a stroke based on thorough follow-up exams.

There are 100 unique patients, each patient was recorded with 4 clinical variables and 100 EEG variables. Only 35 patients have the completed records in clinical variables and EEG variables. Proportion of missing observations in the clinical and EEG variables are all below 5% compared to the total observations Age(1%), Last known well(3%), Race(2%), E1-E100(4%). Since the missing proportion of the variables itself is relatively small, it has little effect on the corresponding variable, we assumed that the missing data mechanism is missing completely at random (MCAR). We use median of each variable to fill in missing values instead of omit it. Therefore data analysis is conducted as a complete fashion.

2.2 Scientific Goals

This study mainly focus on the following three Goals:

- Examines the association between RACE and stroke given all other clinical variables .
- It has further been hypnotized that the association between RACE and stroke varies by age. Conduct an appropriate analysis to assess this hypothesis.
- Using both EEG and clinical variables, build a model for pre-hospital stroke prediction. Evaluate the performance of your model and compare it to a simpler model that uses the clinical variables only.

2.3 Statistical Methods

To address the first question, we use a generalized linear model(Model 1) to explore the association between RACE and stroke given all clinical variables. Y_i =1 represents the patient i did have stroke, otherwise, Y_i =0. Let π_i denoted as the probability of patient i having stroke,where $Y_i|X_i \sim$ Bin(1,π_i).

$$logit(\pi_i) = \beta_0 + \beta_1 Igender_i + \beta_2 LKW_i + \beta_3 Age_i + \beta_4 Race_i \tag{1}$$

Where $gender_i$=1 represents the patient is male. Further, we compute robust (sandwich) variance estimate of the coefficients to examine the association between Race and stroke given all other variable.

As for the second scientific goal, it has been documented by[1] that stroke risk varies by age. Stroke risk increases with age, but strokes can—and do—occur at any age. Based on the records from[1] ,64% of people hospitalized for stroke were older than 65 years old. Here, we first recode the age variable into an indicator old, so that patient who are younger 65 years old both coded as 0, and who are older than 65 years are both coded as 1.

the model is

$$logit(\pi_i) = \beta_0 + \beta_1 Igender_i + \beta_2 LKW_i + \beta_3 Iold_i + \beta_4 Race_i \tag{2}$$

To assess the hypothesis that the association between RACE and stroke varies by age, we add the interaction term $RACE * Iold_i$ into the model.

$$logit(\pi_i) = \beta_0 + \beta_1 Igender_i + \beta_2 LKW_i + \beta_3 Iold_i + \beta_4 Race_i + \beta_5 Race_i * Iold_i \tag{3}$$

Test for significance of interaction term was conducted with the likelihood ratio test.The LR statistic was the difference in deviances (reduced model and full model), which follows an approximate chi squared distribution with degrees of freedom equal to the difference in the number of parameters.The null hypotheses is H_0 $\beta_5 = 0$ and the alternative is that the model with interaction term fits better.The significance level the test was taken to be 0.05.

To build a predictive model for stroke, we first split all data into a training set(70% of observations) and a validation set. With training data, considering all covariate main effects inclusion in the model, since the dimensions of the predictors are larger than the dimensions of observations and some electrodes are spatially close to each other, we use the Least Absolute Shrinkage and Selection Operator(LASSO) to control the complexity of binary regression model.

The objective function for the penalized logistic regression is

$$minimize RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

We use the L1 penalty $\sum_{j=1}^{p} |\beta_j|$ here, a large enough λ with L1 penalty, some of the coefficients could become exactly zero(i.e., become excluded from the model)

In this study, Glmnet package in R is used to achieve the objective. Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso. The algorithm is extremely fast, and can exploit sparsity in the input matrix x.

We first applied 10-fold cross-validation to determine the optimal λ for prediction of stroke. Then we explore the interaction between variables in the selected subset, transformation of variables according to diagnostic analysis. A goodness of fit test is also presented for the predictive model. The predictive accuracy of the model was estimated by application to the validation set, and overall performance of the model in the cross-validated training set and test set was summarized with ROC curves.

3 Result

3.1 Exploratory Data Analysis

In this part we perform a preliminary descriptive data analysis to explore the dataset and provided evidence of the difference of clinical variable compared the cases of non stroke and stroke.

To begin with, the histogram of continuous clinical variables(Age,LKW,RACE) was drawn to understand how it is distributed. As it can be observed from figure 1(in Appendix), they are approximately normally distributed. Also, we notice from Table 1, there are only slightly difference between gender and Last know well time between having stroke or not.The older patients are more likely to be diagnosed as stroke and patients with higher RACE are more possible having stroke.

Covariate Median (IQR) or N (%)	Non Stroke (58, 58%)	Stroke (42, 42%)
Age	63 (61, 67)	65 (63, 68)
Male	30 (52%)	22 (52%)
LKW	11 (7, 18)	10 (6.6, 16)
Race	1 (0, 1)	2 (1, 4)

Table 1: Variables stratified by Diagnostic.

Moreover,in order to explore the association between RACE and stroke,we create a plot in R that summarizes the relationship between the outcome variable stroke, the two age group(55-65 years old and 65-77 years old), and the RACE.

As shown in Figure 1, it appears that in both age groups, the proportion of getting stroke increase as RACE increased , with little difference between the age 55-65 group and age 65-77 group in the pattern of change over RACE. There are some fluctuations among RACE score 0-2 , which is possible to be accounted for the unbalanced observations at each RACE score.

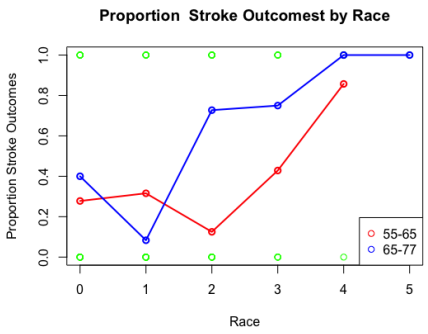

Figure 1: Proportion Stroke vs RACE in age 55-65 and age 65-77

Since we have measured 100 EEG signals which could used to identify patients with acute stroke, we want to explore the difference between this 100 EEG signals among the patients with stroke or not. We compute the mean of EEG signals of the stroke patients and non-stroke patients then plot it.

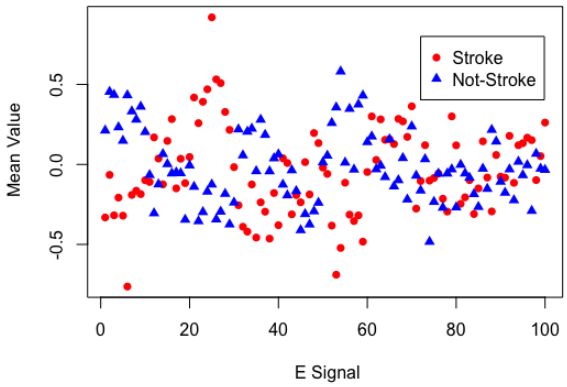

Figure 2: Mean EEG of Stroke VS Non-stroke

As shown in Figure 2, the electroencephalogram pattern of stroke is quite different from the non-stroke's pattern. The summary of E1-E20 signals of non-stroke's patients tends to be larger than E1-E20 of stroke's patients, while E20-E40 signals, the values of stroke's patients tends to larger than non-stroke's patients. After signals 60, the signal change difference between this two groups are relatively small.

3.2 Inference

Modeling stroke risk by RACE To analyze the probability of having stroke associated with RACE, a binomial regression model was constructed adjusting for the all other clinical variables: Age,Gender,LKW.

The result of fitting the main effects model are provided in table , including 95% robust confidence intervals of odds ratio estimates for each coefficients.

	exp(Est)	robust ci95.lo	robust ci95.hi	robust z value	robust Pr(>|z|)
(Intercept)	0.00	0.00	0.13	-2.66	0.01
GenderM	1.02	0.41	2.52	0.04	0.97
LKW	0.97	0.90	1.05	-0.77	0.44
Age	1.11	1.02	1.21	2.37	0.02
RACE	1.86	1.29	2.68	3.33	0.00

Table 2: Result of model 1.

As we can observe from table, the estimated odds of having stroke for RACE is 1.86(95% robust CI:(1.29, 2.68), p value$< 10^{-3}$) . We construct a Likelihood ratio test to test the association of RACE with the probability of having Stroke . The result of the test statistic is 12.921 with p-value 0.00032. There is statistically significant evidence that RACE is associated with a higher probability of having stroke. For the population of patients, 1 unit increases in RACE, the estimated odds of patients having stroke are 86% (95% robust CI:(1.29, 2.68) higher than the original RACE given similar age, gender and LKW.

Hypothesis about RACE and stroke varies by age As the previous model and [1] shown age is positively associated with the probability of having stroke. Age is a confounder here, it is causally related with the predictor of interest (RACE) and the outcome of interest (stroke). To have a better interpretation, we recode the age, I(old)

is the indicator of age>65(classified as higher risk patients). Then we fit the binomial regression model with the interaction term.

	exp(Est)	robust ci95.lo	robust ci95.hi	robust z value	robust Pr(>|z|)
(Intercept)	0.40	0.13	1.25	-1.57	0.12
GenderM	1.06	0.42	2.69	0.13	0.89
LKW	0.97	0.90	1.05	-0.73	0.46
RACE	1.53	1.00	2.34	1.97	0.05
old	0.43	0.05	3.55	-0.78	0.43
RACE:old	2.22	0.80	6.13	1.54	0.12

Table 3: Result of model 3

From the table 3, we found that the estimated odds of patients older than 65 years old having stroke are 122% (95% robust CI:(0.8, 6.13) higher than the patients that are younger than 65 years old given similar RACE, gender and LKW for 1 unit increases in RACE in population level.

Considering hypothesis that the association between RACE and stroke may differ in age, we test the significance of this effect using the likelihood ratio test. The null hypothesis is there is no difference of the association between RACE and stroke across the age.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	95	118.67			
2	94	115.35	1	3.31	0.0688

Table 4: Result of Likelihood ratio test

The result of the test statistic is 3.31 with p-value 0.0688. There is not enough statistical evidence to support that the association of age and stroke varies by age under a=0.05 levels. That is, the association between RACE and stroke is roughly the same regardless of age.

3.3 Predictive Modeling

For construction and assessment of the predictive model, we consider all data set with all clinical and EEG variables in the model. We first randomly divided the observations into a training set consisting of 70% of the observations, with the remaining observations placed in the test set for later validation of the predictive model.

The predictors included in the model were selected via Least Absolute Shrinkage and Selection Operator(LASSO). To obtain the optimal λ, we use misclassification error as the criterion for 10-fold cross-validation. We plot the result of cross-validation

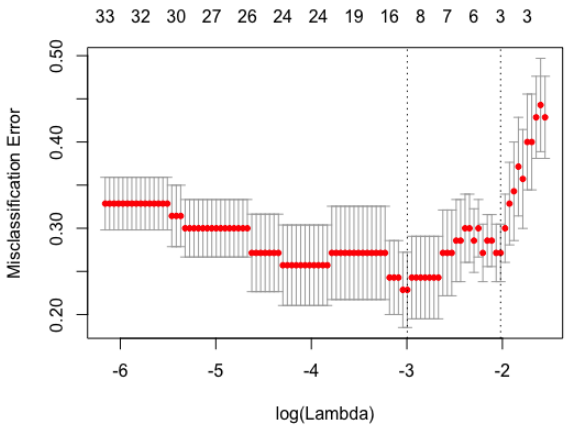

Figure 3: Result of 10-fold cross validation

The optimal λ we achieved via 10-fold cross validation is 0.05001193. In the LASSO binomial regression with optimal λ, most of the predictors coefficients are shrinkage to 0 under penalty , besides Age, RACE, E6, E25, E36, E53. That is subset of the predictors selected by LASSO.

Then we fit the binomial regression model with Age, RACE, E6, E25, E36, E53.

$$logit(\pi_i) = \beta_0 + \beta_1 Age_i + \beta_2 RACE_i + \beta_3 E6_i + \beta_4 E25_i + \beta_5 E36_i + \beta_6 E53_i \tag{5}$$

Their estimates are shown as:

	exp(Est)	robust ci95.lo	robust ci95.hi	robust z value	robust Pr(>|z|)
(Intercept)	0.00	0.00	0.16	-2.18	0.03
Age	1.25	0.97	1.61	1.70	0.09
RACE	4.69	1.44	15.32	2.56	0.01
E6	0.25	0.10	0.62	-2.95	0.00
E25	3.64	1.06	12.55	2.04	0.04
E36	0.31	0.13	0.75	-2.59	0.01
E53	0.39	0.23	0.65	-3.62	0.00

There is statistically significant evidence shown that RACE, E6, E25, E36, E53 are associated with the stroke under a=0.05 levels.

Further exploration of interaction terms indicates no more significant contribution to AIC(Table 5 in Appendix), and diagnostics of linear form implying no need for variable transformation(Figure 8 in Appendix). A Hosmer-Lemeshow goodness of fit test with P-value 0.89 shows adequate fit to the data(Table 6 in Appendix).

Comparison of predictive model and simpler model First, to compare with simpler model that uses the clinical variables only, we conduct the likelihood ratio test to compare the simpler model with our predictive model. The null hypothesis is the simpler model fits good, the alternative hypothesis is our predictive model fits good.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	67	78.49			
2	63	33.72	4	44.78	0.0000

The likelihood ratio statistics is 44.78 with p value $< 10^{-3}$. There is significant statistical evidence shown that our predictive model fits better.

Figure 5 and 6 show the ROC curve for the predictive model applied to training and test sets. The AUCs of the predictive model are 96.8% and 89.8% respectively . The AUCs of the simpler model that uses the clinical variables only are 75.5% and 68.1% respectively.

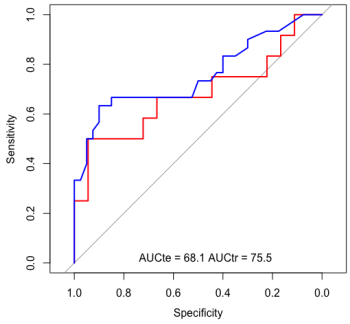

Figure 4: AUC of simple model

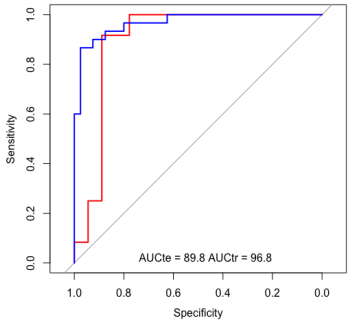

Figure 5: AUC of predictive model

The summary result of Confusion matrix of the simpler model(only with clinical variables) and predictive model is:

	Simple model	Predictive model
Sensitivity	0.6667	0.8333
Specificity	0.6667	0.9167
Pos Pred Value	0.7500	0.9375
Neg Pred Value	0.5714	0.7857
Prevalence	0.6000	0.6000
Detection Rate	0.4000	0.5000
Detection Prevalence	0.5333	0.5333
Balanced Accuracy	0.6667	0.8750

Since specificity measures the proportion of non-stroke (Y=0) that are correctly identified. Sensitivity measures the proportion of stroke (Y=1) that are correctly identified. We prefer higher sensitivity here since it is more important to confirmed diagnosis of the patients having stroke(Y=1). We found that the predictive model has 17% sensitivity higher than the simpler model that only included clinical variables.

4 Discussion

In our study, we analyze the binomial data, 100 observations with 105 variables. We first explore the association between the RACE and stroke given all other clinical variables. And it is found that there is statistically significant evidence of an association between RACE and having stroke such that patients who have 1 unit higher RACE compared to the patients given similar age, gender and Last Known Well Time are at a 86% increased risk of having stroke.

We conduct a likelihood ratio analysis to assess the hypothesized that the association between RACE and stroke varies by age. We added the interaction term into the model, and compared the deviance between same effect model with and without interaction term. The result of the test statistic is 3.31 with p-value 0.0688.There is not enough statistical evidence to support that the association of age and stroke varies by age under a=0.05 levels. That is, the association between RACE and stroke is roughly the same regardless of age.

Predictive model regards age, RACE, E6, E25, E36, E53 as stronger predictors. So in general we could say patients with higher RACE, larger E25 and smaller values of E6, E36, E53 signals may be more likely to be diagnosed as stroke. The predictive model includes clinical and EEG variables(age, RACE, E6, E25, E36, E53) performs much better than the simple model that only includes the clinical variables(age RAE).The AUCs of

predictive model are 0.213 and 0.217 higher than the simple model with only clinical variables applied to training set and test set.

There are some limitations of the analysis. Firstly, we only have 100 observations at a local hospital and we are not told whether the observations were chosen randomly. In addition, all observations are age around 55-77 years old. Lack of observations for younger patients. The inference we made on the population level had been limited, we only could draw inference about this hospital with the paients age around 55-77years old. Moreover, to better evaluate the predictive model, we should collect the data from other hospitals then assess the predictive ability of our predictive model with all data.

Secondly, a more thorough analysis of the structure of missing data is required to validate the above conclusions.

Thirdly, some references[3][5][9] indicated that there is also race difference and regions differences on the stroke risk. White man are more likely to have stroke compared with others. Eastern American residents are more likely to have stroke compared with Western American residents. Moreover, some cardiovascular risk factors such as smoke, BMI are also associated with the stroke. Hence in further study, researchers should collect more observations and variables from different hospitals.

5 Reference

[1]Kertesz, A., Sheppard, A. N. N. (1981). The epidemiology of aphasic and cognitive impairment in stroke: age, sex, aphasia type and laterality differences. Brain: a journal of neurology, 104(Pt 1), 117-128.

[2]Party, I. S. W. (2012). National clinical guideline for stroke (Vol. 20083). London: Royal College of Physicians.

[3]Bassetti, C. L., Milanova, M., Gugger, M. (2006). Sleep-disordered breathing and acute ischemic stroke: diagnosis, risk factors, treatment, evolution, and long-term clinical outcome. Stroke, 37(4), 967-972.

[4]Persson, M., Fhager, A., Trefná, H. D., Yu, Y., McKelvey, T., Pegenius, G., ... Elam, M. (2014). Microwave-based stroke diagnosis making global prehospital thrombolytic treatment possible. IEEE Transactions on Biomedical Engineering, 61(11), 2806-2817.

[5]Schwamm, L. H., Reeves, M. J., Pan, W., Smith, E. E., Frankel, M. R., Olson, D., ... Fonarow, G. C. (2010). Race/ethnicity, quality of care, and outcomes in ischemic stroke. Circulation, 121(13), 1492.

[6]Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[7]Jordan, K. G. (2004). Emergency EEG and continuous EEG monitoring in acute ischemic stroke. Journal of Clinical Neurophysiology, 21(5), 341-352.

[8]Nuwer, M. R., Jordan, S. E., Ahn, S. S. (1987). Evaluation of stroke using EEG frequency analysis and topographic mapping. Neurology, 37(7), 1153-1153.

[9]Ottenbacher, K. J., Smith, P. M., Illig, S. B., Linn, R. T., Fiedler, R. C., Granger, C. V. (2001). Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. Journal of clinical epidemiology, 54(11), 1159-1165.

6 Appendix

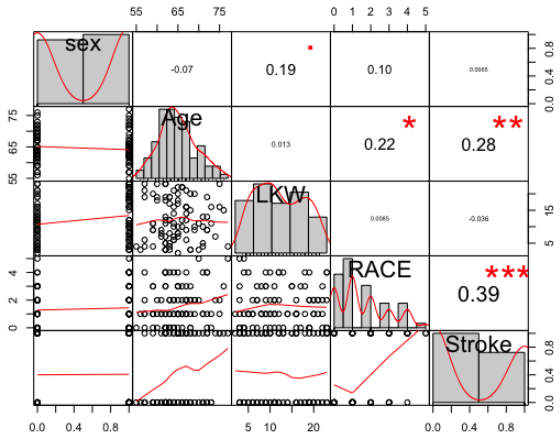

Figure 6: Correlation among clinical variables

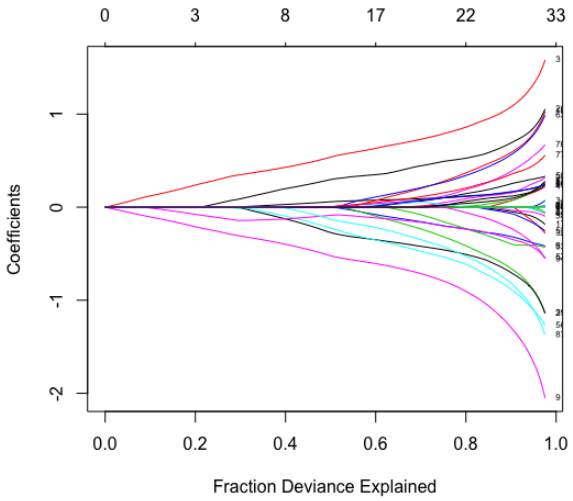

Figure 7: LASSO coefficients

Model	AIC
no interaction	47.72
RACE :age	49.68
RACE :E6	48.63
RACE :E25	51.73
RACE :E36	46.65
RACE :E53	49.60

Table 5: Result of AIC

	chisqstat	df	pVal
1	3.56	8.00	0.89

Table 6: Result of GOF

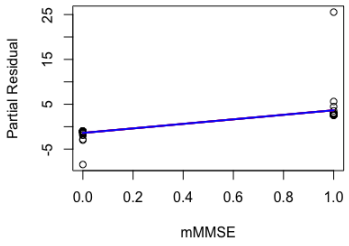

Figure 8: RACE and mMMSE

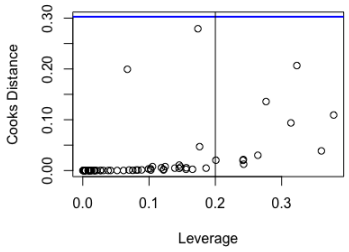

Figure 9: Cook's distance to examine influential points

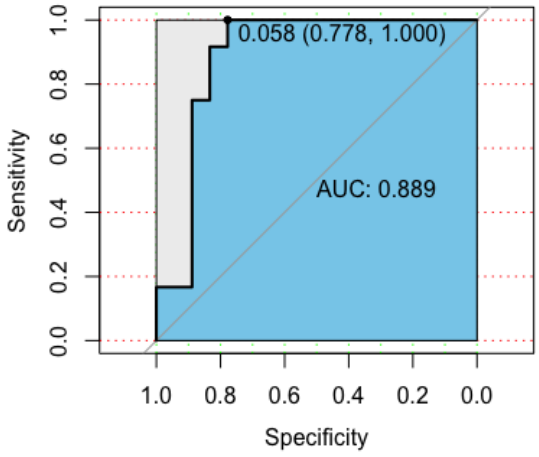

Figure 10: AUC of predictive model